DOCUMENT RESUME

ED 398 274 TM 025 468

AUTHOR Lau, Che-Ming Allen; And Others
TITLE Using Unidimensional IRT Models for Dichotomous Classification via Computerized Classification Testing with Multidimensional Data.
PUB DATE Apr 96
NOTE 30p.; Paper presented at the Annual Meeting of the American Educational Research Association (New York, NY, April 8-12, 1996).
PUB TYPE Reports - Evaluative/Feasibility (142) -- Speeches/Conference Papers (150)

EDRS PRICE MF01 Plus Postage. PC Not Available from EDRS.
DESCRIPTORS *Classification; *Computer Assisted Testing; *Educational Testing; *Item Response Theory; Monte Carlo Methods; *Robustness (Statistics); Simulation
IDENTIFIERS Dichotomous Variables; *Multidimensionality (Tests); One Parameter Model; Three Parameter Model; *Unidimensionality (Tests)

ABSTRACT

This study focused on the robustness of unidimensional item response theory (UIRT) models in computerized classification testing against violation of the unidimensionality assumption. The study addressed whether UIRT models remain acceptable under various testing conditions and dimensionality strengths. Monte Carlo simulation techniques were used to assess the robustness of the UIRT models by comparing the false positive classification rates, false negative classification rates, and numbers of items administered for classification or mastery under the different conditions. In this study, both unidimensional three-parameter and one-parameter models were found robust with the sequential probability ratio testing procedure in computerized classification testing. The UIRT model chosen, the strength of the dimensionality, and testing conditions had an impact in terms of classification accuracy and test efficiency. (Contains 2 figures, 4 tables, and 11 references.) (Author/SLD)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)
☑ This document has been reproduced as received from the person or organization originating it.
☐ Minor changes have been made to improve reproduction quality.
• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

PERMISSION TO REPRODUCE AND DISSEMINATE THIS MATERIAL HAS BEEN GRANTED BY
CHE-MING ALLEN LAU
TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

ED 398 274

Using Unidimensional IRT Models for Dichotomous Classification via Computerized Classification Testing with Multidimensional Data[1]

Che-Ming Allen Lau
Abdel-fattah A. Abdel-fattah
Judith A. Spray

ACT, 1996

[1] The opinions expressed in this article are those of the authors and should not be mistaken to represent the policy of the ACT.

BEST COPY AVAILABLE

TM025468

Abstract

This study focuses on the robustness of unidimensional item response theory (UIRT) models in computerized classification testing (CCT) against violation of the unidimensionality assumption. This study addresses whether the UIRT models are still acceptable under various testing conditions and dimensionality strength.

Monte Carlo simulation techniques were used to assess the robustness of the UIRT models by comparing the false positive classification rates, false negative classification rates, and numbers of items administered for classification or mastery under the different conditions.

In this study, both unidimensional 3 parameter and 1 parameter models (UIRT 3-PL & UIRT 1-PL) were found robust with sequential probability ratio testing procedure in computerized classification testing. The UIRT model chosen, the strength of the dimensionality, and testing conditions had an impact in terms of classification accuracy and test efficiency.

Using Unidimensional IRT Models for Dichotomous Classification via Computerized Classification Testing with Multidimensional Data

Computerized classification testing (CCT) items are usually calibrated with UIRT models rather than multidimensional IRT (MIRT) models due to availability of the unidimensional software for item calibration and simplicity of the unidimensional models. If the unidimensionality assumption is violated, there could be some adverse effect on the pass/fail classification in CCT. In that case, it is important to find out if CCTs with UIRT models are still useful for making a dichotomous classification decision.

It was found that the UIRT 3 parameter logistic model (UIRT 3-PL) could be used for calibrating 2 dimensional item data and the item parameter estimates could be used for CCT under a variety of conditions related to the strength of the dimensionality. In other words, under certain conditions, the use of UIRT 3-PL item parameter estimates for CCT gave nominal classification errors when the true, underlying model was MIRT 3-PL (Abdel-fattah, Lau, & Spray, 1995). In the current study, the interest was in comparing the UIRT 1 parameter logistic model (UIRT 1-PL) to the UIRT 3-PL under comparable conditions. The justification for using the UIRT

1-PL model was that this model is quite popular for many certification and licensure computerized testing programs.

Review

Unidimensional IRT Models

The UIRT 3-PL model defines the probability of a correct response as

$$p_i(\theta_j) = p_i(x_{ij} = 1 \mid \theta_j) = c_i + \frac{1 - c_i}{1 + \exp[-Da_i(\theta_j - b_i)]}$$

where

$p_i(\theta_j)$ is the probability that an examinee with ability θ_j answers item i correctly,

x_{ij} is the dichotomous (0/1) response to item i by person j,

a_i is the item i discrimination parameter,

b_i is the item i difficulty parameter,

c_i is the item i "guessing" parameter,

D is a scaling factor applied to make the logistic model as close as possible to the normal ogive model, in general, D=1.7.

The UIRT 1-PL model is a special cases of the UIRT 3-PL model. For the UIRT 1-PL model, the discrimination parameter

is constant across all items and the "guessing" parameter is set equal to zero for all items.

In CCT, items are selected to be administered according to the item information either at the previous estimated θ or at the cutting point: the greater the information, the greater the probability of selection. Information at a given ability level varies directly as the square of the item discriminating power, a_i (Lord, 1980). The item information function, $I\{\theta, u_i\}$ can be stated as follows (Lord, 1980):

$$I\{\theta, u_i\} = \frac{P_i'^2}{P_i Q_i}$$

where

$P_i = P_i(\theta)$ is the item response function,

$Q_i = 1 - P_i$, and

P_i' is the derivative of P_i with respect to θ.

Item Selection and Administration

In the context of computerized classification testing, items may be selected based on different criteria, depending on the particular procedure being applied. Two procedures, sequential Bayes (SB) and sequential probability ratio testing (SPRT), are frequently used (e.g., Reckase, 1983; Kingsbury & Weiss, 1983).

With the sequential Bayes procedure, the first item(s) for administration is/are usually of average difficulty because the ability level of the examinee is unknown. After obtaining the first estimate of ability, items with the greatest information at this estimation of the examinee's θ are selected. With the sequential probability ratio testing procedure, items with the greatest information at the cut point are selected. Spray and Reckase (in press) administered items that were maximally informative at the cut point and found that the SPRT procedure was more efficient than the SB procedure. The SPRT procedure required fewer items to achieve the same level of classification accuracy as the sequential Bayes procedure.

Mastery Decision and Test Termination Rule

The mastery decision is based on the cut score. The basic idea is as follows: if the examinee's θ estimate is at or above the cut score, then the examinee is classified as a master; otherwise, the examinee is classified as a nonmaster. Because the scores or abilities of the examinees are estimates, the degree of accuracy of these estimates is usually considered.

Two kinds of errors can happen in such dichotomous classification systems: 1) classifying an examinee whose θ

below θ_c as mastery; and 2) to classify the examinee whose θ equals or above θ_c as nonmastery. The following notation is used to represent these two types of errors:

α = prob(classify as mastery | $\theta < \theta_c$)

β = prob(classify as nonmastery | $\theta \geq \theta_c$)

where

θ_c is the cut point

Generally, two criteria are imposed to terminate the CCT administration. First, if the degree of accuracy of estimation reaches a certain preset degree, a mastery/nonmastery decision can be made and the test administration stops. Second, if the maximum permitted number of items has been reached, the test administration stops.

Sequential Probability Ratio Testing Procedure

The sequential probability ratio testing (SPRT) developed by Wald (1947) can be used to classify examinees into two categories (mastery/nonmastery). Reckase (1983) noted that the SPRT procedure could be modified and applied to adaptive mastery testing with UIRT models. SPRT theory does not require the probabilities to be constant from item to item within the pool (Spray and Reckase, 1987).

With the SPRT procedure, the test items are selected to be administered to each examinee according to the amount of information at the cut point. The test items are ranked at the cut point with respect to the amount of information provided. That is, the greater the information, the greater the priority of administration. The use of the SPRT procedure in computerized mastery testing has not been widespread. The general procedure of SPRT is outlined below (Reckase, 1983).

Suppose an examinee has ability θ_j. The decision about the examinee's status (pass or fail) is made on the basis of a consideration of two simple hypotheses:

$$H_0: \theta_j = \theta_0$$

vs.

$$H_1: \theta_j = \theta_1$$

where

θ_j is an unknown parameter, and

θ_0 and θ_1 represent decision points that correspond to lower and upper limits, respectively, of the passing criterion, δ, where $\theta_0 < \delta < \theta_1$.

If $P(\theta_j)$ is the probability that examinee j responds correctly to an item, and $Q(\theta_j) = 1 - P(\theta_j)$ is the

BEST COPY AVAILABLE

probability that examinee i responds incorrectly to an item, then,

$$\pi(\theta_j) = \text{Prob}(X=x \mid \theta=\theta_j) = P(\theta_j)^x Q(\theta_j)^{1-x},$$

where

x = 1, correct response, and

x = 0, incorrect response.

The functions, $\pi(\theta_1)$ and $\pi(\theta_0)$, are called likelihood functions of x, and a ratio of these two functions, $L(x) = \pi(\theta_1) / \pi(\theta_0)$, is called a likelihood ratio and

$$L(x_1, x_2, \ldots, x_n \mid \theta_0, \theta_1) = \frac{\pi_1(\theta_1)\pi_2(\theta_1)\ldots\pi_n(\theta_1)}{\pi_1(\theta_0)\pi_2(\theta_0)\ldots\pi_n(\theta_0)}$$

The likelihood ratio is compared to the boundaries, A and B.

where

$A = (1-\beta)/\alpha$, and

$B = \beta/(1-\alpha)$

where α and β are the error probabilities defined as follows:

Prob(choosing H_1 | H_0 is true) = α (false positive)

Prob(choosing H_0 | H_0 is true) = 1 - α (correct decision)

Prob(choosing H_0 | H_1 is true) = β (false negative)

Prob(choosing H_1 | H_1 is true) = 1 - β (correct decision)

The mastery decision and test termination rules in SPRT are as follow:

If $L(x_1,x_2,\ldots,x_n \mid \theta_0,\theta_1) \geq A$, then H_1 is accepted. In this case, the examinee is classified as a master and the test administration is stopped.

If $L(x_1,x_2,\ldots,x_n \mid \theta_0,\theta_1) \leq B$, then H_0 is accepted. In this case, the examinee is classified as a nonmaster and the test administration is stopped.

If $B < L(x_1,x_2,\ldots,x_n \mid \theta_0,\theta_1) < A$, then no decision is made and another item is selected to be administered.

The region from θ_0 to θ_1 is the indifference region. The distance, $|\theta_0 - \theta_1|$ is the width of the indifference region. Test length is a function of this region.

Compensatory Multidimensional Item Response Theory Model

Mckinley and Reckase (1983) proposed a compensatory multidimensional IRT (COMIRT) 3 parameter model (COMIRT 3-PL) that permits high ability on one dimension to compensate for low ability on the others. This model defines the probability of a correct response as

$$P_i(x_{ij} = 1 \mid \underline{a}_i, di, c_i; \underline{\theta}_j) = c_i + \frac{1 - c_i}{1 + \exp\left[-D\left(\sum_{k=1}^{m} a_{ik}\theta_{jk} + d_i\right)\right]}$$

where

x_{ij} is the dichotomous (0/1) response to item i by person j,

$\underline{a}_i = (a_{i1}, a_{i2}, \ldots, a_{im})$ is the vector of item discrimination parameters for item i,

d_i is a scale parameter that is related to the difficulty parameter for item i,

c_i is the guessing parameter for item i,

D is a scaling factor applied to make the logistic model as close as possible to normal ogive model, in general, D=1.7,

$\underline{\theta}_j = (\underline{\theta}_{j1}, \underline{\theta}_{j2}, \ldots, \underline{\theta}_{jm})$ is the vector of ability parameters for person j, and

m is the number of dimensions.

Correlation between Theta 1 and Theta 2

Ansley and Forsyth (1985) used two-dimensional item sets with noncompensatory MIRT model and noted that as the correlation between the ability dimensions increased, the average absolute differences between values of the true ability parameters and the estimated ability parameters decreased.

Folk and Green (1989) used two-dimensional simulated item sets with compensatory MIRT model in both adaptive and nonadaptive testing format to study the effects of the correlation between thetas in item parameter estimates. The UIRT 3-PL model was used to do the calibration. They found that as the correlation between ability dimensions decreased, the estimated a-parameter increasingly emphasized either a_1 or a_2. The average estimated a-parameter values were highest with when the correlation between theta 1 and theta 2 was 1. They suggested that if the data did not deviate greatly from unidimensionality, the UIRT model can provide a reasonable approximation.

Using simulated 2-dimensional data with both compensatory and noncompensatory MIRT models and different correlation between the abilities, Ackerman (1987) found that as the correlation between the 2-dimensional abilities increased, the response data appeared to become more unidimensional.

From these studies, it could be concluded that as the correlation between thetas in the 2-dimensional case increases, the response data "acts" more like a unidimensional data set.

BEST COPY AVAILABLE

Research Questions

This study will address the following research questions.

1. Does computerized mastery testing using the sequential probability ratio testing procedure based on UIRT models have acceptable accuracy in terms of type I and type II error rates when the true IRT models underlying the responses are two dimensional?
2. Is the accuracy of the UIRT model-based procedures affected by the use of different UIRT models such as the 3-parameter model or the 1-parameter model?
3. Is the accuracy of the UIRT model-based procedures affected by the correlation between the two sets of ability parameters?
4. Is the accuracy of the UIRT model-based procedures affected by the following testing conditions:
   a) range of test length and
   b) level of the cut score?

Methodology

The primary purpose of this study is to examine whether the unidimensional item response theory models used in making mastery decisions with the sequential probability ratio test (SPRT) procedure produce acceptable accuracy in

terms of type I errors (false positive) and type II errors (false negative) when the assumption of unidimensionality is violated.

Simulation Procedure

The basic idea of the simulation used in this study is to generate 2 dimensional dichotomous data (i.e., 1/0) and then calibrate the items with UIRT models under different conditions. The robustness of UIRT models is then investigated by comparing the false positive classification rates, false negative classification rates, and numbers of items administered for classification or mastery under the different conditions.

Research Design

This study was a 2x4x2x2 design with 32 conditions investigated. The conditions are listed below:

1) UIRT 3-PL versus UIRT 1-PL models
2) correlation between theta 1 and theta 2 (.00, .30, .60, .90)
3) passing score (.4, .8)
4) CCT testing constraints required for classification (min=15, max=50; min=1, max=360)

BEST COPY AVAILABLE

Data Source

In this study, the original 2 dimensional parameters (a_1, a_2, d & c) were based on actual test data. The item pool was created from an ACT Assessment Math Usage Test form. The pool consisted of 360 multiple-choice items that required students to use their reasoning skill to solve practical problems in mathematics. 2,000 examinee responses were used to calibrate the items with the computer program NOHARM, (Fraser, 1983) in which compensatory MIRT model was applied.

From the estimated, 2 dimensional item parameters, dichotomous scores were generated with the compensatory MIRT model, from two bivariate distributions of theta 1 and theta 2. The correlation between theta 1 and theta 2 was set to .00, .30, .60, and .90.

Data Analysis

With the 4 sets of simulated 1/0 responses, the computer program BILOG was used to calibrate these 2 dimensional dichotomous scores data with UIRT 3-PL model and UIRT 1-PL model. One thousand randomly drawn respondents were used for the item calibrations. The estimated item parameters from UIRT 3-PL and UIRT 1-PL models were obtained and used for the mastery tests.

In order to distinguish between true pass and true fail in the true model, a threshold function line, $f(\theta_1, \theta_2)$, is defined as:

$$f(\theta_1, \theta_2) = \frac{1}{360} \sum_{i=1}^{360} p_i(x_{ij} = 1 \mid a_{i1}, a_{i2}, d_i, c_i; \theta_{j1}, \theta_{j2}) \text{ - cutting point}$$

where

cutting point = 0.4, and 0.8.

Then, the true parameters, together with the threshold function, $f(\theta_1, \theta_2)$ of 2-D COMIRT 3-PL model, and the false parameters from the UIRT model, together with the corresponding theta mapped from $p(\theta)$= 0.4 or 0.8 of UIRT (either UIRT 3-PL or UIRT 1-PL) will be used as input to the computer program MIRTSPRT created for this study. MIRTSPRT was used to calculate the average type I error rates, type II error rates, and number of items used to make the mastery decision under different combinations of conditions for comparing 2-D COMIRT 3-PL model with UIRT 1-PL or UIRT 3-PL models. Figure 1 and 2 summarize these procedures.

The sequential probability ratio testing procedure will be used to determine the pass/fail status using the UIRT model.

The mastery decision based on the true model (i.e., 2-D COMIRT 3-PL) is defined as:

BEST COPY AVAILABLE

If $f(\theta_1,\theta_2) \geq 0$, the examinee truly passes.

If $f(\theta_1,\theta_2) < 0$, the examinee truly fails.

The decision (pass or fail) made by the UIRT models are then compared to the decision based on the true model. If the decision based on true item parameters and the corresponding θ_1 and θ_2 is consistent with the decision based on false (i.e., either UIRT 3-PL or UIRT 1-PL) test parameters and the corresponding theta, there will be no error. However, if the decision based on true item parameters is to fail the examinee while the decision based on UIRT item parameters is to pass the examinee, a type I error (false positive) occurs. If the decision based on true item parameters is to pass the examinee while the decision based on UIRT item parameters is to fail the examinee, a type II error (false negative) occurs.

The fixed factors for all simulations of this design are listed below:

1. the number of dimensions (2)
2. the number of items in the pool (360)
3. the means and standard deviations of the generated thetas
4. the means and standard deviations of a_1, a_2, d, & c in 2-D COMIRT 3-PL model
5. the number of simulation replications (300)

6. the width of the SPRT indifference region ($|\theta_0 - \theta_1| = 0.5$) (i.e., $\theta_0 = \delta - 0.25$, $\theta_1 = \delta + 0.25$, where δ is the passing criterion)
7. the nominal error rates, α and β in the SPRT procedure (0.05)

Result

Table 1 through table 4 summarized the results. Table 1 and table 2 indicated the error rates and average test length for investigated conditions. Table 3 recorded the accumulated information for different conditions and table 4 showed the corresponding thetas estimated by UIRT models.

UIRT 3-PL Model Verse UIRT 1-PL Model

The test lengths of UIRT 3-PL model were found consistently shorter than that of UIRT 1-PL model. The accumulated information value based on UIRT model was consistently higher in 3-PL model than that of 1-PL model.

When the test length constraint was applied (i.e., min=15, max=50), UIRT 3-PL model made fewer classification errors than UIRT 1-PL model. However, where there was no test length constraint condition (i.e., min=1, max=360), the results reversed.

BEST COPY AVAILABLE

Effect of Changing Correlation between Theta 1 and Theta 2.

It was found that as the correlation between thetas became higher, fewer items were needed to make the classification in all conditions.

With the UIRT 3-PL model, the higher the correlation, the smaller the error when the passing score was set to .4. With the UIRT 1-PL model, the higher the correlation, the smaller the error when the passing was set to .4 as the test length constraint was applied.

Passing Score Setting

When the passing score was set to .8, fewer errors were found and fewer items were needed to make the decision throughout all conditions.

For the 3-PL model, more accumulated information was gained when the cut score was .8. For 1-PL model, more accumulated information was gained when the cut score was .4 and when maximum test length was 360; the accumulated information was about the same when maximum test length was 360.

When the passing score was set to .4, type I error occurred consistently more often than type II error throughout all conditions. When the passing score was set to .8, type II errors were higher than type I error when correlation between theta 1 and theta 2 was less or equal to

.6 throughout all conditions. It can be implied that: when the passing score was set to .8, the estimated corresponding thetas had a "type I error bias"; when the passing score was set to .4, the estimated corresponding thetas had a "type II error bias".

Effect of Applying Test Length Constraints

The UIRT 3-PL model gained similar classification precision under either test length constraint condition or no test length constraint application. Without the test length constraint, the UIRT 3-PL model used fewer items to make the classification.

The UIRT 1-PL model made fewer errors but used more items when there was no length constraint. However, the UIRT 1-PL model made more error, and used fewer items when there was a length constraint.

When the test length constraint was applied, the UIRT 3-PL model has smaller total error rates than the UIRT 1-PL model. However, if no test length constraint was applied, the UIRT 1-PL model made smaller total error rates than the UIRT 3-PL model.

Discussion

The SPRT procedure was found robust and useful for mastery decision even when the unidimensionality assumption

BEST COPY AVAILABLE

was violated. As the function base on true model departed more from the cut-point, fewer classification errors occurred and fewer items were needed to make the mastery decision.

When the underlying data set was 2 dimensional, the UIRT 3-PL model performed more efficiently than the UIRT 1-PL model. The UIRT 1-PL model consumed about 1.5 times to 3 times as many items as UIRT 3-PL model to make the mastery decision. Therefore, the UIRT-1PL model could be still useful if more items were allowed to be administered. (e.g., by setting the maximum number of items to be higher)

When the true data set was 2 dimensional but calibrated by the UIRT model, a "passing theta bias" occurred. That is: for some passing points, the passing theta was underestimated and more type I error occurred; for other passing point, the passing theta was overestimated and more type II error occurred. It seemed that the passing theta bias had an interaction effect with the correlation of the thetas.

Test length constraint is commonly applied in CCT. Consideration should be included which UIRT model is adopted in addition to content balance and security.

Even though the true data was 2 dimensional, the information function based on UIRT models is still a useful index for selecting items to administer in the CCT.

In this study, a trend was found that, as the correlation between the two thetas increased, the error rates decreased and the number of items used to make the mastery classification decreased. That implied that as the correlation between thetas increased, the data set "behaved" more unidimensional which was consistent with previous studies. (e.g., Abdel-fattah, Lau, & Spray, 1995).

References

Abdel-fattah, A. A., Lau, C. A., & Spray, J. A. (1995). The effect of model misspecification on classification decisions made using a computerized test: MIRT versus UIRT. Paper presented at the meeting of the Psychometric Society, Minneapolis.

Ackerman, T. (1987). A comparison study of the unidimensional IRT estimation of compensatory and noncompensatory multidimensional item response data (ACT Research Report Series 87-12). Iowa City, IA: American College Testing.

Ackerman, T. (1991). The use of unidimensional parameter estimates of multidimensional items in adaptive testing. Applied Psychological Measurement, 15(1), 13-24.

Ansley, T. N., & Forsyth, R. A. (1985). An examination of the characteristics of unidimensional IRT parameter estimates derived from two-dimensional data. Applied Psychological Measurement, 9, 37-48.

Folk, V. G., & Green, B. F. (1989). Adaptive estimation when the unidimensionality assumption of IRT is violated. Applied Psychological Measurement, 13(4), 373-389.

BEST COPY AVAILABLE

Kingsbury, G. G., & Weiss, D. J. (1983). A comparison of IRT-based adaptive mastery testing and a sequential mastery testing procedure. In D. J. Weiss (Ed.), New horizons in testing; latent trait test theory and computerized adaptive testing. (pp. 257-283) New York: Academic Press.

Lord, F. M. (1980). Applications of item response theory to practical testing problems. Hillsdale, NJ: Erlbaum.

Reckase, M. D. (1983). A procedure for decision making using tailored testing, In D. J. Weiss (Ed.), New horizons in testing; latent trait test theory and computerized adaptive testing (pp. 237-255). New York: Academic Press.

Spray, J. A., Reckase, M. D.(in press). Comparison of SPRT and Sequential Bayes procedures for classifying examinees using an adaptive test. Journal of Educational and Behavioral Statistics.

Spray, J., Reckase, M. D. (1987). The effect of item parameter estimation error on decisions made using the sequential probability ratio test. (ACT Research Report Series 87-1). Iowa City, IA: American College Testing.

Wald, A. (1947). Sequential Analysis. New York: Dover Publications, Inc.

Table 1. Error Rates and Average Test Length for Test Length: min=15, max=50

		UIRT 3-PL				UIRT 1-PL			
$r_{\theta:\theta_2}$	cut score	false positive	false negative	total error	ATL	false positive	false negative	total error	ATL
.00	.4	.0357	.0245	.0602	21.69	.0441	.0393	.0834	35.41
.00	.8	.0064	.0140	.0204	15.85	.0030	.0190	.0221	22.46
.30	.4	.0311	.0196	.0507	19.90	.0531	.0231	.0762	32.08
.30	.8	.0071	.0143	.0214	15.84	.0019	.0246	.0265	21.05
.60	.4	.0245	.0162	.0407	18.81	.0485	.0224	.0709	29.35
.60	.8	.0082	.0140	.0222	15.73	.0017	.0278	.0295	19.93
.90	.4	.0212	.0111	.0323	17.86	.0518	.0108	.0626	27.32
.90	.8	.0106	.0057	.0163	15.58	.0017	.0189	.0206	19.05

Note: ATL is the average test length; $r_{\theta:\theta_2}$ is the value of correlation between two thetas.

Table 2. Error Rates and Average Test Length for Test Length: min=1, max=360

		UIRT 3-PL				UIRT 1-PL			
$r_{\theta_1\theta_2}$	cut score	false positive	false negative	total error	ATL	false positive	false negative	total error	ATL
.00	.4	.0302	.0224	.0506	26.58	.0270	.0155	.0425	61.06
.00	.8	.0067	.0135	.0202	6.14	.0016	.0121	.0137	27.86
.30	.4	.0249	.0166	.0415	18.41	.0265	.0140	.0405	49.45
.30	.8	.0094	.0152	.0246	6.57	.0020	.0158	.0178	25.31
.60	.4	.0251	.0159	.0411	14.03	.0306	.0131	.0437	41.42
.60	.8	.0111	.0138	.0249	4.63	.0012	.0163	.0175	22.41
.90	.4	.0214	.0156	.0370	11.23	.0292	.0089	.0381	34.37
.90	.8	.0168	.0070	.0238	4.35	.0018	.0076	.0094	19.58

Note: ATL is the average test length; $r_{\theta_1\theta_2}$ is the value of correlation between two thetas.

Table 3. Accumulated Information at Each Condition

		accumulated information			
		UIRT-3PL		UIRT-1PL	
$r_{\theta_1\theta_2}$	cut score	max=50	max=360	max=50	max=360
.00	.4	27.9562	64.2432	10.1665	59.6267
.00	.8	63.3480	141.1518	9.8029	41.5512
.30	.4	37.2529	83.7545	13.2217	77.8227
.30	.8	73.7314	166.9306	12.7553	54.0386
.60	.4	45.1929	100.8731	16.4185	96.7734
.60	.8	88.9715	197.3844	15.8277	67.0963
.90	.4	55.2764	120.1003	19.7152	116.4226
.90	.8	114.8079	241.5028	18.9875	80.3927

Table 4. Cutting scores and the Corresponding Cutting Thetas

		corresponding θ	
$r_{\theta_1\theta_2}$	cut score	UIRT 3-PL	UIRT 1-PL
.00	.4	-.1600	-.3560
.00	.8	1.6360	1.9640
.30	.4	-.1480	-.3320
.30	.8	1.4640	1.6960
.60	.4	-.1320	-.3200
.60	.8	1.3520	1.5000
.90	.4	-.1160	-.3120
.90	.8	1.2480	1.3520

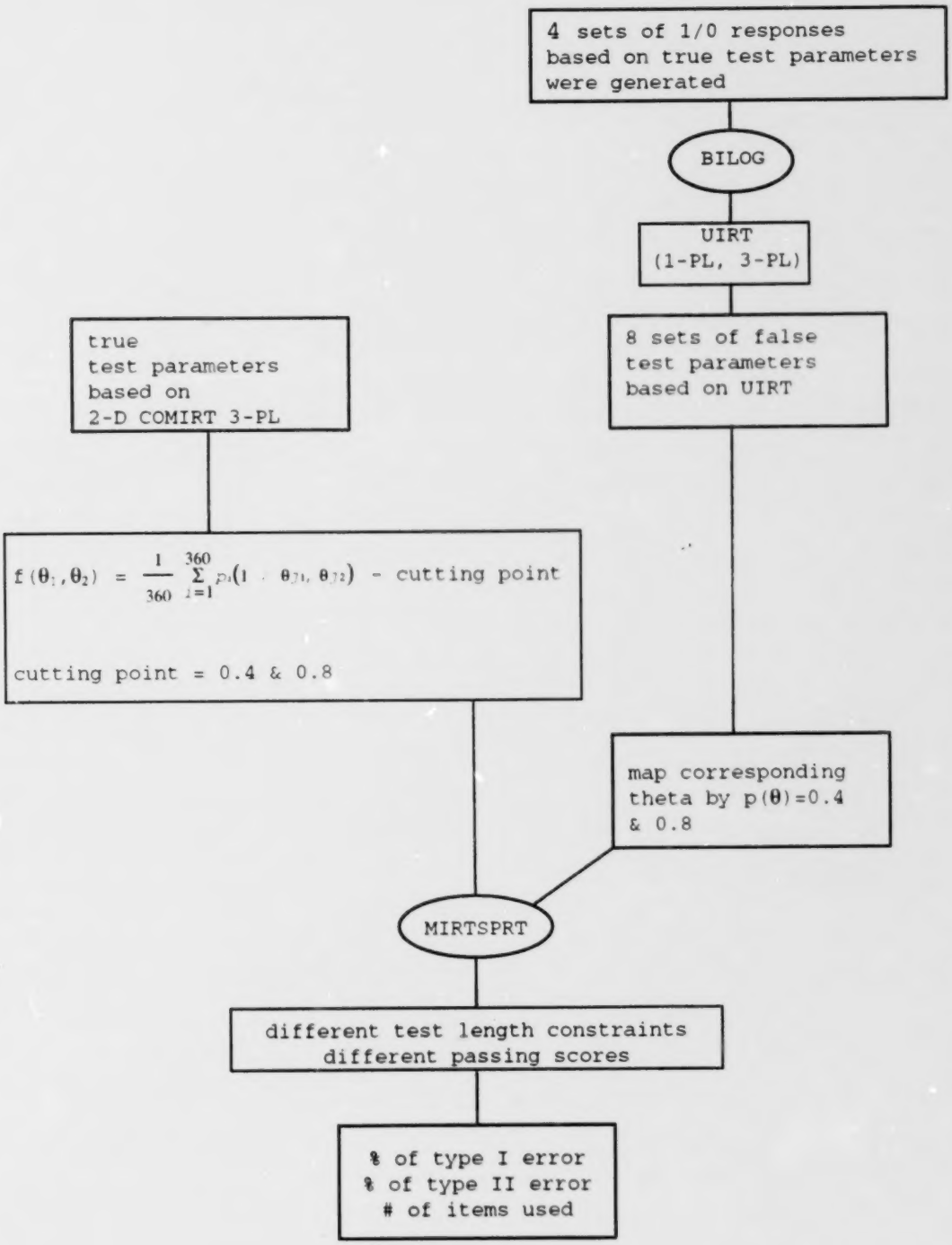
4 sets of 1/0 responses based on true test parameters were generated
BILOG
UIRT (1-PL, 3-PL)
8 sets of false test parameters based on UIRT
true test parameters based on 2-D COMIRT 3-PL
$f(\theta_1, \theta_2) = \frac{1}{360} \sum_{i=1}^{360} p_i(1 - \theta_{j1}, \theta_{j2})$ - cutting point
cutting point = 0.4 & 0.8
map corresponding theta by p(θ)=0.4 & 0.8
MIRTSPRT
different test length constraints
different passing scores
% of type I error
% of type II error
# of items used

Figure 1. Data Analysis Flow Chart

BEST COPY AVAILABLE

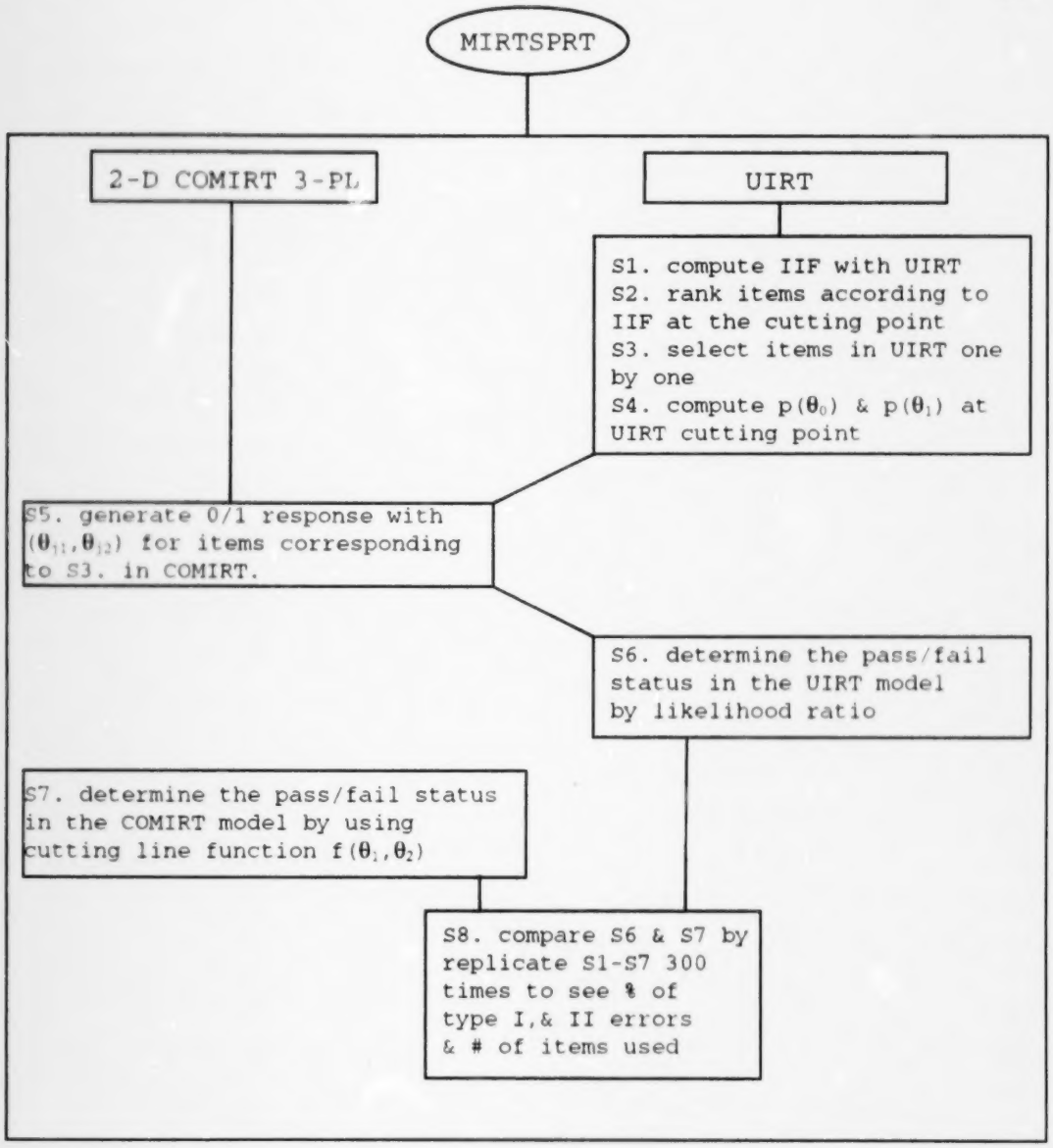
MIRTSPRT
2-D COMIRT 3-PL
UIRT
S1. compute IIF with UIRT
S2. rank items according to IIF at the cutting point
S3. select items in UIRT one by one
S4. compute p(θ0) & p(θ1) at UIRT cutting point
S5. generate 0/1 response with (θj1,θj2) for items corresponding to S3. in COMIRT.
S6. determine the pass/fail status in the UIRT model by likelihood ratio
S7. determine the pass/fail status in the COMIRT model by using cutting line function f(θ1,θ2)
S8. compare S6 & S7 by replicate S1-S7 300 times to see % of type I,& II errors & # of items used

Figure 2. Type I and Type II Errors Calculation

Note: IIF is the item information function; S1-S8 are step 1 to step 8